Tổng hợp dữ liệu

Nội dung

- Hiểu các xử lý cơ bản trên tập kết quả đã giải chuẩn
- Các phương pháp tổng hợp dữ liệu
- Liên hệ các lựa chọn khác nhau để tổng hợp dữ liệu với các loại thang đo khác nhau
- Xây dựng các bảng tổng hợp, các bảng cross-tab, pivot sử dụng các xử lý trên tập kết quả và các lựa chọn tổng hợp

Cắt (slicing) và xắt (dicing)

- Thuật ngữ được sử dụng cho việc chọn một tập con các hàng trong một bảng đa chiều gọi là cắt (slicing): loại bỏ các hàng mà chúng ta không quan tâm. Ví dụ về quản lý bán hàng: chỉ chọn các dữ liệu giao dịch bán hàng máy tính,...
- Thuật ngữ sử dụng cho việc chọn một số các cột (chiều) của bảng dữ liệu đa chiều gọi là xắt (dicing): loại bỏ các cột không quan tâm. Ví dụ: chỉ quan tâm đến tên của các nhân viên bán hàng trong một số phân tích hiệu quả bán hàng của nhân viên.
- Với hai hoạt động trên cho phép chúng ta 'cắt và xắt' các bảng đa chiều thành các bảng nhỏ hơn phù hợp hơn với các mục đích cụ thể.

Cắt (slicing) và xắt (dicing)

	Cột 1	Cột 2	Cột 3	Cột 4
Hàng 1	A	C	…	…
Hàng 2	A	D	…	…
Hàng 3	B	C	…	…
Hàng 4	B	D	…	…

Xắt (dicing)

	Cột 1	Cột 2
Hàng 1	A	C
Hàng 2	A	D
Hàng 3	B	C
Hàng 4	B	D

Cắt (slicing)

	Cột 1	Cột 2	Cột 3	Cột 4
Hàng 1	A	C	…	…
Hàng 2	A	D	…	…

Cắt và xắt (Slicing and dicing)

	Cột 1	Cột 2
Hàng 1	A	C
Hàng 2	A	D

Bảng tổng hợp dữ liệu

- Loại xử lý thứ ba và thứ tư trên các bảng liên quan đến tổng hợp dữ liệu, tạo dữ liệu tổng hợp từ dữ liệu đã có.
- Xây dựng một số hàng bằng cách thay thế bằng một hàng với các giá trị tổng hợp đại diện. Ví dụ, chúng ta có bảng dữ liệu các giao dịch bán hàng và co bảng đó lại bằng việc nhóm theo mã đơn hàng hoặc nhóm theo sản phẩm.

Bảng tổng hợp dữ liệu

Don_hang	Khach_hang	San_pham	Gia	So_luong
1	Hương	Cam	1	6
1	Hương	Nến	7	34
2	Loan	Rượu	10	24
2	Loan	Cam	1	4
2	Loan	Táo	1	35

Bảng tổng hợp dữ liệu

Don_hang	Khach_hang	San_pham	Gia	So_luong
1	Hương	Cam	4	40
2	Loan	Táo	4	63

Don_hang	Khach_hang	San_pham	Gia	So_luong
-	-	Cam	1	10
-	-	Nến	7	34
-	-	Rượu	10	24
-	-	Táo	1	35

Bốn hoạt động xử lý trên bảng dữ liệu

- Quá trình gộp dữ liệu thành các giá trị tổng hợp được gọi là cuộn dữ liệu lên (rolling up).
- Quá trình ngược lại mở rộng giá trị dữ liệu tổng hợp thành các giá trị giao dịch gọi là quá trình đào sâu dữ liệu (drilling down).
- Như vậy, có bốn hoạt động xử lý trên bảng dữ liệu được giải chuẩn là cắt, xắt, cuộn lên và đào sâu.
- Các gói phần mềm trí tuệ doanh nghiệp thường hỗ trợ thực hiện các hoạt động xử lý này theo trình tự bất kỳ.

THANG ĐO DỮ LIỆU TỔNG HỢP

- Thang đo danh định: dữ liệu có thể đếm được nhưng không xếp hạng được. Ví dụ, sản phẩm, tên nhân viên kinh doanh, nhóm kinh doanh...
- Thang đo thứ tự: dữ liệu có thể đếm được và các giá trị có thể được xếp hạng theo một trình tự có ý nghĩa, nhưng sự khác biệt giữa các giá trị là không biết được. Ví dụ: lựa chọn 'Rất không đồng ý', 'Hơi không đồng ý', 'Trung lập , 'Hơi đồng ý', 'Rất đồng ý'....➔ Tổng hợp dữ liệu một giá trị này lớn hơn giá trị kia nhưng chúng ta không biết là cao hơn bao nhiêu.

THANG ĐO DỮ LIỆU TỔNG HỢP

- Thang đo khoảng: dữ liệu có thể đếm được, các giá trị có thể xếp hạng được và sự khác biệt bao nhiêu giữa các giá trị là rõ ràng. Ví dụ: nhãn thời gian Time Stamp,.. ➔ Đặc tính xác định của các thang đo khoảng là chúng ta có thể đếm các giá trị, xếp hạng chúng và cộng trừ chúng.
- Thang tỷ lệ: giống thang đo khoảng và thêm vào đó giá trị 0 được xác định phân biệt. Điểm không biểu thị thuộc tính có số lượng bằng không. Ví dụ: lượng bán, giá sản phẩm,...

THANG ĐO TỶ LỆ (tiếp)

- Thang tỷ lệ: Lượng bán tại điểm không có nghĩa là không có sản phẩm nào được bán. Giá tại điểm không nghĩa là không phải trả tiền cho sản phẩm. Sự tồn tại của điểm không là quan trọng vì nó phân biệt giữa thang tỷ lệ và thang khoảng và nó cho phép chúng ta nhân, chia các giá trị dữ liệu mà không thể thực hiện được với thang khoảng.
- Các giá trị tỷ lệ cho phép diễn đạt một giá trị là một phần của giá trị khác ở cùng chiều.
- Thang đo tỷ lệ: các giá trị có thể đếm được, xếp hạng được, cộng trừ được và nhân chia được

Một số nhầm lẫn thường gặp

- Nhầm lẫn giữa thang đo dang định và thang đo thứ tự: Ví dụ, các NVBH là thang đo danh định vì chúng ta có thể đếm được. Nếu phân tích SQL sắp xếp các NV theo thứ tự alphabet của tên nên có thể hiểu nhầm NVBH ở thang đo thứ tự, là ngẫu nhiên theo tên và không phản ánh xếp hạng chất lượng NV.
- Giá trị danh định thường bị nhầm lẫn là có thể sắp xếp được bằng cách sử dụng các phép đo khác. Ví dụ, chúng ta có thể sắp xếp các NVKD theo năng suất bán hàng, nên có thể ở thang đo thứ tự chứ không phải ở thang đo danh định?

Một số nhầm lẫn thường gặp

- Thang đo khoảng thường bị nhầm lẫn với thang đo tỷ lệ ở chỗ sự khác nhau giữa các giá trị khoảng có thể nhân, chia được. Ví dụ, hai lần giai đoạn hai tháng là một giai đoạn bốn tháng do đó sự khác biệt có thể được biểu diễn trong thang đo tỷ lệ.
- Sự khác biệt giữa cách chúng ta xếp hạng dữ liệu trên thang đo thứ tự và cách chúng ta sắp xếp thứ tự khi chúng ta lấy dữ liệu bằng một câu truy vấn từ bảng quan hệ.

Truy vấn dữ liệu

```
SELECT ten
FROM        trang_thai_khach_hang
ORDER BY    ten
```

Thứ tự của đáp ứng sẽ là theo thứ tự alphabet như sau:

ten

==========

Lien_lac

Khach_hang_thuc_su

Khach_hang_tiem_nang

Khach_hang_trien_vong

Truy vấn dữ liệu

```
SELECT      hang, ten
FROM        trang_thai_khach_hang
ORDER BY    hang
```

Sẽ cho kết quả sau:

Hang	ten
1	Lien_lac
2	Khach_hang_tiem_nang
3	Khach_hang_trien_vong
4	Khach_hang_thuc_su

CÁC TÙY CHỌN TỔNG HỢP DỮ LIỆU

- Một giá trị tổng hợp là một giá trị biểu diễn một nhóm các giá trị của cùng một thuộc tính nhưng từ các thể hiện (hàng) khác nhau.
- Ví dụ, chúng ta có các giá trị 3, 5, và 7 là giá của ba sản phẩm, giá trị tổng hợp có thể là 15 biểu diễn tổng giá trị các sản phẩm.
- Giá trị tổng hợp khác có thể cho ví dụ này là 5 biểu diễn giá trị giá trung bình của ba sản phẩm.
- Thang đo của biến là yếu tố quyết định trong lựa chọn tổng hợp dữ liệu.

Các hoạt động tổng hợp dữ liệu thông dụng

- Đếm (count): là tần suất. Phép đo này biểu diễn số lần xảy ra của các giá trị dữ liệu cụ thể trong tập dữ liệu. Tần suất có thể được sử dụng với tất cả các kiểu biến và dùng được với các biến đo danh định.
- Tối thiểu, tối đa (min, max): Cung cấp giá trị thấp nhất và cao nhất trong hạng. Các biến cần phải được điều chỉnh tỷ lệ cho một hạng để có nghĩa do đó các biến đo đanh định không áp dụng được. Các biến ở thang đo còn lại có thể sử dụng lựa chọn này để tổng hợp dữ liệu.

Các hoạt động tổng hợp dữ liệu thông dụng

- Tính tổng: Đây là giá trị tổng khi tất cả các giá trị dữ liệu được tổng lại. Các giá trị dữ liệu có thể được tính toán số học bao gồm tối thiểu là phép cộng trừ các giá trị. Chỉ áp dụng cho các biến đo khoảng và thang đo tỷ lệ.
- Tính trung bình: tổng hợp thông tin về tâm của phân bố các giá trị dữ liệu. Có các phép đo trung bình khác nhau tùy thuộc vào thang đo. Với các biến đo danh định, phép đo tổng hợp phù hợp là mode, là giá trị xảy ra nhiều nhất. Với các biến đo thứ tự, phép đo tổng hợp trung bình là trung vị. Với các biến đo khoảng và tỷ lệ, phép đo là giá trị trung bình.

Các hoạt động tổng hợp dữ liệu thông dụng

- Biến động (variation): cung cấp thông tin về sự dàn trải của dữ liệu xung quanh tâm. Phép đo này không áp dụng cho các biến đo danh định do không tồn tại tâm. Với thang đo thứ tự, khoảng tứ phân vị (IRQ) có thể được xác định. Với các biên đo khoảng và tỷ lệ, phép đo biến động là độ lệch chuẩn (SD), gần hay xa so với giá trị trung bình.
- Dạng (shape): mô tả một cách chi tiết sự dàn trải của dữ liệu quanh tâm. Kurtosis (độ nhọn) và Skewness (độ xoắn) là các phép đo phổ biến nhất cho các biến đo khoảng và tỷ lệ.

Các hoạt động tổng hợp dữ liệu thông dụng

- Kurtosis biểu thị độ phẳng của phân bố dữ liệu: kurtosis càng cao thì dữ liệu càng gần với giá trị trung bình và ngược lại.
- Skewness biểu thị tính đối xứng của dạng. Phép đo độ xoắn xác định có hay không dữ liệu tập trung hơn ở một phía của giá trị trung bình so với phía khác.
- Các phép đo tổng hợp về dạng thường ít được sử dụng hơn so với các phép đo khác.

Thuộc tính dẫn xuất

- Thuộc tính dẫn xuất (derived attibute) là một thuộc tính mà các giá trị của nó là kết hợp theo một cách định nghĩa trước của các giá trị khác của cùng thể hiện.
- Ví dụ: thuộc tính ‘Giá thành’ có thể được tính bằng cách nhân giá trị của thuộc tính ‘Đơn giá’ với thuộc tính ‘Số lượng’.
- Thuộc tính dẫn xuất không được lưu trong tập dữ liệu nhưng do biết cách tính chúng nên có thể tính toán ra chúng khi cần khai thác.

Thuộc tính dẫn xuất

- Thuộc tính dẫn xuất (derived attibute) là một thuộc tính mà các giá trị của nó là kết hợp theo một cách định nghĩa trước của các giá trị khác của cùng thể hiện.
- Ví dụ: thuộc tính ‘Giá thành’ hay ‘Lợi tức’ có thể được tính bằng cách nhân giá trị của thuộc tính ‘Đơn giá’ với thuộc tính ‘Số lượng’.
- Thuộc tính dẫn xuất không được lưu trong tập dữ liệu nhưng do biết cách tính chúng nên có thể tính toán ra chúng khi cần khai thác.

Thuộc tính dẫn xuất

- Giá trị dẫn xuất (in nghiêng) luôn tham chiếu đến một thể hiện và khác với giá trị tổng hợp (in thường) luôn tham chiếu đến nhiều thể hiện dữ liệu.

Don_hang	San_pham	Gia	So_luong	Loi_tuc
1	Cam	1	6	*6*
1	Nến	7	34	*238*
				244

Bảng tóm tắt

- Một bảng tóm tắt cung cấp dữ liệu tổng hợp của một thuộc tính được nhóm lại theo các giá trị của một thuộc tính khác.
- Bảng tóm tắt cung cấp một tập con các cột và hàng và được coi là khung nhìn súc tích của tập kết quả giải chuẩn.
- Các gói phần mềm trí tuệ doanh nghiệp cho phép tóm tắt và đào sâu dữ liệu một cách tương tác.

Bảng tóm tắt

- Minh họa bảng tóm tắt cho biến đo danh định: tổng lợi tức được tóm tắt theo nhóm bán hàng

Nhom_bh	Loi_tuc
Alpha	*1,051*
Beta	*889*
	1,940

- Minh họa bảng tóm tắt cho biến đo khoảng: Tổng lợi tức chia theo tuần.

Tuan	Loi_tuc
1	*244*
2	*692*
3	*260*
4	*424*
4	*324*
	1,940

Bảng tần số

- Trong một bảng tần số, các giá trị của một thuộc tính được nhóm vào một cột và cột thứ hai chỉ thị số lần xảy ra của mỗi giá trị trong tập kết quả gốc.
- Ví dụ cho các biến trong thang đo danh định, là đếm số sản phẩm được bán.

San_pham	Dem
Cam	4
Táo	4
Rượu	4
Xà phòng	4
Nến	4
	20

Bảng tần số

- Các bảng tần suất có thể được xây dựng cho các biến khoảng và tỷ lệ: các ngăn sẽ được xác định.
- Các ngăn được cắt từ các phần của thang đo sao cho chúng biểu diễn toàn bộ khoảng ở đó các biến được đo.
- Mỗi ngăn biểu diễn một phần bằng nhau của thang đo.
- Khi các ngăn được xây dựng, các giá trị dữ liệu nằm trong một ngăn có thể được đếm số lần xuất hiện.

Bảng cross-tab

- Một bảng cross-tab là một bảng tóm tắt cho hai chiều.
- Giá trị của bảng cross-tab cung cấp cái nhìn sâu sắc trong mối quan hệ giữa hai biến trong đó cross-tab chia nhỏ số tổng.
- Ví dụ: cung cấp thông tin tổng quan về lợi tức kinh doanh của mỗi nhóm bán hàng mỗi tháng

Bảng cross-tab

- Bảng cross-tab lợi tức chia theo nhân viên kinh doanh theo tuần
- Nhân viên tên Nam ở nhóm Alpha rất thành công ở tuần 1 và tuần 2 và không thành công ở các tuần sau

Nhan_vien	Nhom_bh	Tuan_1	Tuan_2	Tuan_3	Tuan_4	Tuan_5	Tong
Nam	Alpha	244	75	0	0	0	319
Hương	Alpha	0	279	29	424	0	732
Tuấn	Beta	0	108	231	0	294	633
Thủy	Beta	0	230	0	0	26	256
		244	692	260	424	320	1,940

Bảng cross-tab

- Bảng cross-tab của sản phẩm được bán chia theo lợi tức theo sản phẩm được

Loi_tuc	Cam	Táo	Rượu	Xà phòng	Nến	Tong
0-50	4	4	0	1	1	10
51-100	0	0	0	3	0	3
101-150	0	0	0	0	0	0
151-200	0	0	3	0	0	3
201-250	0	0	1	0	3	4
	4	4	4	4	4	120

Bảng cross-tab

- Các hàng và cột trong bảng cross-tab có ý nghĩa khác với các hàng và cột trong bảng quan hệ hoặc tập kết quả SQL.
- Trong bảng quan hệ, các cột được coi là các thuộc tính của các thực thể và các hàng là các thể hiện của các thực thể.
- Trong bảng cross-tab, cả hàng và cột là các giá trị tổng hợp.

Bảng pivot

- Một bảng pivot là phiên bản tương tác của bảng cross-tab.
- Bảng pivot cho phép tạo các cross-tab một cách tương tác.
- Để tạo bảng pivot, cần khai thác các bảng dữ liệu gốc.
- Sử dụng bảng pivot, các cấu trúc tóm tắt có thể được thay đổi bằng cách đổi chỗ các hàng và các cột mà chúng ta thấy phù hợp sử dụng phương pháp kéo - nhả tương tác.

XIN CẢM ƠN!